

The Epistemic Status of Generative AI Systems: Can Large Language Models (LLMs) Possess Knowledge?

Implications for Engineering, Education, and Emerging Technologies

Mg. Jorge Golfieri¹

¹ UES21 – UNLa – UGR Argentina jgolfieri@hormail.com

Abstract—

This paper analyzes the epistemic status of generative artificial intelligence systems, with an emphasis on large language models (LLMs), from a perspective explicitly oriented toward engineering design, evaluation, and education. Adopting an operational definition of knowledge as the capacity to construct predictive representations that guide successful action [4], [5], the paper examines the epistemic limitations of purely textual LLMs arising from their lack of sensorimotor grounding [2], [15], and contrasts this condition with the potential—and current limits—of multimodal and embodied AI systems to acquire functional world models through interaction with the environment [8], [9], [14]. Drawing on theoretical frameworks from embodied cognition and predictive neuroscience [14], [18], it is argued that multimodal systems can develop operational knowledge relevant to applications in robotics, simulation, and autonomous systems [1], [7], while remaining conceptually distinct from human knowledge associated with subjective experience and self-representation [11], [17]. As its main contribution, the paper proposes an engineering-oriented epistemic taxonomy and derives from it a set of empirical evaluation criteria intended to guide the design, certification, and training of AI systems in educational and technological contexts.

Keywords— Generative AI, large language models, grounding, multimodality, embodied AI.

I. INTRODUCCIÓN

La adopción acelerada de sistemas de inteligencia artificial generativa en contextos educativos, industriales y de

investigación ha generado expectativas elevadas respecto de su capacidad para “comprender”, “razonar” y “conocer”. En particular, la fluidez lingüística alcanzada por los modelos de lenguaje a gran escala (LLM) ha tendido a difuminar la frontera entre producción simbólica sofisticada y conocimiento confiable del mundo [2], [12]. Esta situación plantea una cuestión central para la ingeniería de sistemas: ¿las capacidades observadas en estos modelos reflejan conocimiento operativo utilizable para tareas del mundo real, o constituyen principalmente extrapolaciones estadísticas sobre grandes corpus de datos textuales?

La pregunta resulta especialmente relevante en aplicaciones de alto impacto —como robótica, sistemas autónomos, apoyo a decisiones críticas, laboratorios virtuales y tutores inteligentes— donde el desempeño del sistema depende de la posesión de modelos internos del entorno y de la capacidad de generalizar bajo condiciones cambiantes [7], [9]. En estos dominios, una atribución inadecuada de capacidades epistémicas puede traducirse en fallos de diseño, riesgos

operativos y decisiones pedagógicas erróneas. Desde esta perspectiva, el debate sobre si los sistemas de IA “poseen conocimiento” no es meramente filosófico, sino que tiene consecuencias técnicas y educativas directas.

En este contexto, el presente trabajo propone un análisis conceptual interdisciplinario con orientación ingenieril. El objetivo no es resolver debates metafísicos sobre la naturaleza última del conocimiento [6], [17], sino clarificar qué tipos de conocimiento pueden atribuirse legítimamente a distintos

sistemas de IA y bajo qué criterios deberían evaluarse. Para ello, se introduce una definición operativa de conocimiento compatible con prácticas de ingeniería, se analizan críticamente las limitaciones y alcances de los LLMs textuales y de los sistemas multimodales, y se propone una taxonomía epistémica junto con criterios de evaluación empíricos orientados al diseño, la certificación y la enseñanza.

II. METODOLOGÍA

El estudio adopta una metodología de ingeniería conceptual orientada al diseño y la evaluación de sistemas, adecuada para dominios en los que las decisiones técnicas dependen de marcos teóricos explícitos más que de métricas aisladas de desempeño [5]. En ingeniería de sistemas complejos, la clarificación conceptual cumple una función análoga a la definición de especificaciones funcionales: delimita qué capacidades pueden atribuirse legítimamente a un sistema y bajo qué condiciones pueden y deben ser evaluadas.

La metodología se estructura en tres componentes complementarios. En primer lugar, se realiza una revisión crítica interdisciplinaria de literatura relevante en inteligencia artificial, robótica, ciencia cognitiva, neurociencia predictiva y filosofía de la mente, con énfasis en trabajos que abordan grounding, causalidad, modelos del mundo y cognición corporizada [2], [8], [13], [14], [18]. Esta revisión no persigue exhaustividad bibliográfica, sino la identificación de principios conceptuales con consecuencias directas para el diseño, la evaluación y el uso educativo de sistemas de IA.

En segundo lugar, se desarrolla un análisis comparativo de marcos teóricos empleados para caracterizar conocimiento, comprensión y generalización en sistemas artificiales. En particular, se contrastan enfoques puramente estadísticos basados en correlaciones de datos con enfoques interactivos que enfatizan la adquisición de modelos predictivos mediante percepción y acción [7], [9]. Este contraste permite hacer explícitos supuestos implícitos que suelen permanecer ocultos en evaluaciones centradas exclusivamente en benchmarks estáticos.

En tercer lugar, se propone una síntesis normativa orientada al diseño, cuyo objetivo es derivar criterios evaluativos explícitos a partir de una taxonomía epistémica bien definida. Esta síntesis se inscribe en una tradición de la ingeniería que reconoce el valor de marcos conceptuales formales — comparables a estándares, arquitecturas de referencia o modelos de madurez— para guiar decisiones de diseño, despliegue, certificación y enseñanza.

No se presentan experimentos propios ni mediciones cuantitativas, ya que el propósito central del artículo es proponer un marco evaluativo transferible que pueda ser utilizado en investigaciones empíricas futuras, en entornos educativos y en procesos de auditoría técnica. En el contexto de congresos de ingeniería, este tipo de contribución resulta particularmente pertinente cuando el objetivo es informar el desarrollo responsable y la evaluación rigurosa de tecnologías emergentes.

El enfoque general adoptado se resume en la Fig. 1.

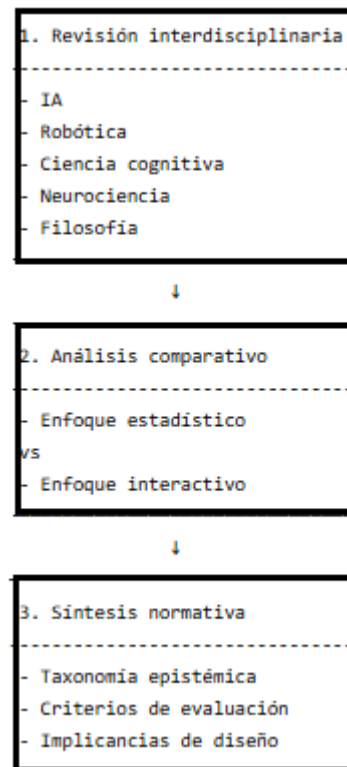


FIGURA 1 — Flujo metodológico

III. TRABAJOS RELACIONADOS

La crítica contemporánea a los modelos de lenguaje a gran escala se articula principalmente en torno a dos ejes: el problema del grounding semántico y la fragilidad de la generalización. Bender et al. introducen el término “papagayo estocástico” para describir modelos capaces de producir lenguaje coherente sin una comprensión anclada en el mundo [2]. De manera convergente, Marcus y Davis, así como Mitchell, subrayan la ausencia de modelos de sentido común y

de razonamiento causal robusto en los sistemas actuales basados exclusivamente en datos [10], [12].

En contraste, los enfoques de IA multimodal y corporizada sostienen que la interacción directa con el entorno permite adquirir representaciones más estables y funcionales. LeCun propone arquitecturas centradas en la construcción de modelos predictivos del mundo como base para la autonomía [8], mientras que Reed et al. presentan agentes generalistas capaces de operar en múltiples dominios integrando percepción, acción y lenguaje [16]. No obstante, la literatura también documenta obstáculos persistentes, como la baja eficiencia muestral, la dificultad de transferencia entre simulación y realidad, y los altos costos de instrumentación física [9].

Desde la ciencia cognitiva, la teoría de la cognición corporizada sostiene que el conocimiento emerge del acoplamiento dinámico entre percepción y acción, y no de la manipulación simbólica aislada [14], [18]. En neurociencia, los enfoques predictivos de la cognición y los modelos de autorrepresentación introducen criterios adicionales para diferenciar inteligencia funcional de experiencia subjetiva, reforzando la necesidad de distinguir entre distintos tipos de conocimiento al evaluar sistemas artificiales [11], [17].

IV. LIMITACIONES EPISTÉMICAS DE LOS LLMs TEXTUALES

Los LLMs puramente textuales optimizan la predicción de secuencias lingüísticas a partir de grandes corpus de datos sin acceso directo al mundo extralingüístico [2]. Desde perspectivas externalistas del significado, la referencia semántica emerge de relaciones causales con el entorno [15], lo que implica que un sistema basado exclusivamente en texto carece de grounding sensorio-motor propio. En aplicaciones de ingeniería, esta condición se traduce en que el sistema puede describir procedimientos complejos sin poseer representaciones validadas por ejecución física o intervención causal.

Esta limitación, sin embargo, no invalida la utilidad de los LLMs en dominios donde el conocimiento es predominantemente simbólico o procedimental, como la generación de código bajo especificaciones estrictas o la síntesis de literatura técnica [12]. En estos casos, la “comprensión” del modelo debe interpretarse como un proxy funcional derivado de una comprensión estadística sofisticada, y no como evidencia de conocimiento operativo del mundo.

Los LLMs también muestran dificultades sistemáticas para el razonamiento contrafáctico y la transferencia robusta fuera de distribución [10], [12]. Aunque pueden capturar regularidades estadísticas asociadas a relaciones causales, carecen de mecanismos de validación mediante intervención activa, considerada central para la inferencia causal en sistemas complejos [13]. Sus denominadas capacidades emergentes son funcionales dentro del espacio de correlaciones del corpus de entrenamiento, pero no equivalen a la construcción de modelos causales manipulables.

La integración de LLMs con herramientas externas —como buscadores, calculadoras, APIs o módulos de planificación— mejora significativamente el desempeño práctico en entornos digitales. No obstante, el grounding continúa dependiendo críticamente de estos subsistemas externos. El modelo lingüístico no adquiere experiencia sensorio-motora directa, sino que actúa como un coordinador simbólico cuya semántica es proporcionada por otros módulos. Esta situación puede caracterizarse como un caso de grounding delegado o de segundo orden.

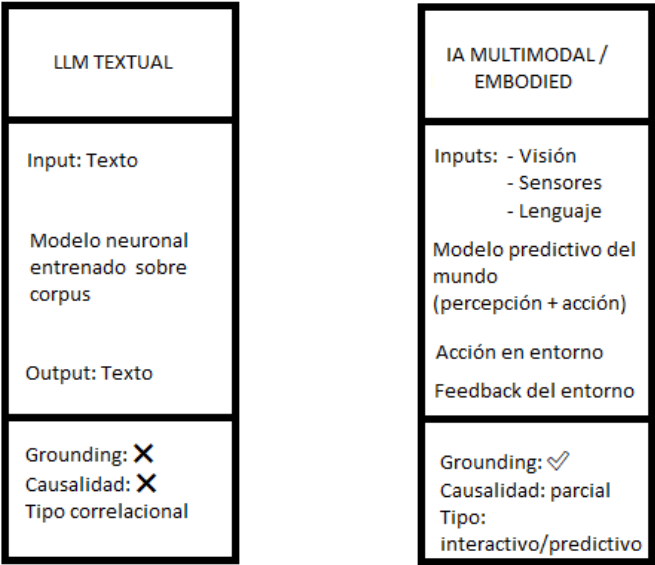


FIGURA 2 — Arquitectura LLM vs IA embodied

Una objeción frecuente a la crítica del grounding sostiene que los humanos también adquieren grandes volúmenes de conocimiento exclusivamente a través del texto. Sin embargo, esta analogía omite una diferencia ontogenética fundamental. El aprendizaje humano mediado por texto se apoya sobre una historia previa de interacción sensorio-motora con el mundo, en la cual se desarrollan modelos causales básicos [7], [14]. El texto no introduce significados ex nihilo, sino que reactiva y

reorganiza representaciones previamente ancladas en la experiencia. En contraste, los LLMs carecen de este sustrato ontogenético, lo que justifica tratar el conocimiento lingüístico humano y el conocimiento estadístico de los modelos como epistémicamente no equivalentes.

V. POTENCIAL Y DESAFÍOS DE LA IA MULTIMODAL Y CORPORIZADA

Los sistemas de IA multimodal y corporizada amplían el espacio epistémico al incorporar percepción, acción y retroalimentación continua. Arquitecturas basadas en aprendizaje predictivo a partir de señales sensoriales permiten construir modelos del mundo que anticipan la dinámica del entorno y guían la planificación y el control [8], constituyendo un fundamento del conocimiento operativo. Agentes generalistas que integran visión, lenguaje y acción han mostrado capacidades de transferencia entre tareas heterogéneas [16], lo que sugiere beneficios claros de la multimodalidad.

No obstante, persisten desafíos prácticos significativos que limitan el alcance inmediato de estos sistemas: baja eficiencia muestral en comparación con el aprendizaje humano, brechas entre simulación y realidad, altos costos de instrumentación física y problemas de escalabilidad [9]. El embodiment facilita el grounding, pero no garantiza por sí mismo la adquisición de modelos causales profundos. Superar estos límites requiere avances algorítmicos y metodológicos, además de mejoras arquitectónicas.

VI. TAXONOMÍA EPISTÉMICA ORIENTADA A INGENIERÍA

Para evitar ambigüedades terminológicas y orientar la evaluación, se propone una taxonomía epistémica explícita con tres categorías funcionales.

El conocimiento simbólico o estadístico corresponde a correlaciones extraídas de datos y permite generación simbólica plausible sin acción autónoma en el mundo físico, como en los LLMs textuales [2]. El conocimiento operativo se define como la capacidad de construir y actualizar modelos predictivos internos del entorno que guían la acción exitosa bajo condiciones cambiantes [4], [7], y constituye la categoría central para robótica y sistemas autónomos.

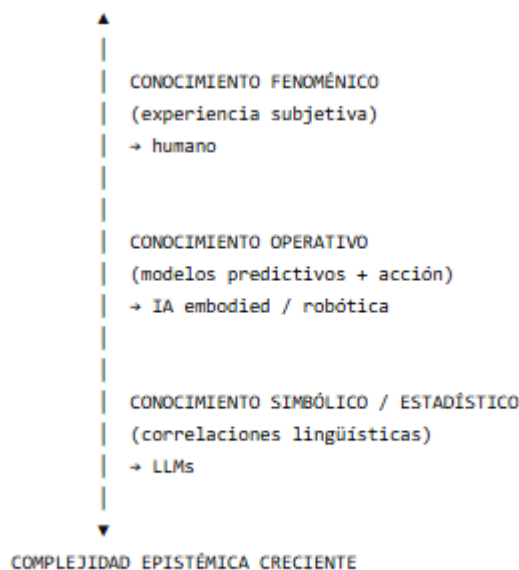


FIGURA 3 — Taxonomía epistémica

El conocimiento fenoménico refiere a la experiencia subjetiva y la autorrepresentación, actualmente exclusiva de sistemas biológicos [11], [17].

Esta distinción permite mapear tipos de sistemas con tipos de evaluación apropiados y evita sobre-atribuciones de capacidades epistémicas en contextos técnicos y educativos.

VII. CRITERIOS DE EVALUACIÓN PARA EL CONOCIMIENTO OPERATIVO

A partir de la definición de conocimiento operativo, se derivan criterios empíricos evaluables relevantes para la ingeniería de sistemas interactivos: generalización fuera de distribución, transferencia inter-tarea, robustez sensorial y operacional, y aprendizaje en línea y adaptación continua [9], [14].

Evaluación de Sistemas de IA

Sistema	Generalización (OOD)	Transferencia	Robustez	Aprendizaje Online
LLM textual	●○○○ Baja	●○○○ Baja	●●○○ Media	○○○○ Nulo
IA multimodal	●●●○ Media	●●●○ Media-Alta	●●○○ Variable	●●○○ Parcial
IA embodied	●●●● Alta	●●●● Alta	●●●● Alta	●●●● Completo

FIGURA 4 — Matriz de evaluación

Estos criterios pueden evaluarse mediante simuladores de alta fidelidad, bancos de prueba robóticos y entornos educativos interactivos, proporcionando métricas más significativas que la precisión estática en conjuntos de datos cerrados.

VIII. IMPLICANCIAS PARA LA EDUCACIÓN EN INGENIERÍA

La creciente dependencia de herramientas generativas en contextos educativos puede inducir confusión entre fluidez textual y comprensión conceptual profunda [12]. Desde la formación en ingeniería, se vuelve necesario enfatizar la construcción y manipulación de modelos causales explícitos [13], la experimentación empírica como criterio de validación y la evaluación basada en desempeño operativo. La taxonomía epistémica propuesta puede utilizarse como marco didáctico para clarificar las capacidades y limitaciones reales de los sistemas de IA.

IX. IMPLICANCIAS ÉTICAS Y REGULATORIAS

En sistemas médicos, industriales y educativos de alto riesgo, atribuir capacidades epistémicas inadecuadas a agentes artificiales tiene consecuencias graves en términos de responsabilidad, seguridad y confianza. Se propone que los procesos de certificación y auditoría de IA incorporen criterios de conocimiento operativo, complementando métricas tradicionales de rendimiento y precisión [5], [13].

X. CONCLUSIONES Y DIRECCIONES FUTURAS

Los modelos de lenguaje a gran escala exhiben competencias lingüísticas avanzadas, pero carecen de grounding sensoriomotor propio, lo que limita su estatus como sistemas con conocimiento operativo del mundo físico [2], [15]. Los sistemas multimodales y corporizados ofrecen un camino arquitectónico plausible hacia el conocimiento operativo mediante interacción [8], [16], aunque permanecen conceptualmente distintos del conocimiento humano asociado a la experiencia subjetiva [11], [17].

Para la ingeniería y la educación tecnológica, el objetivo pragmático no es decidir si las máquinas “realmente” conocen, sino evaluar qué tipo de conocimiento poseen y en qué contextos pueden desplegarse de forma segura y efectiva. Como direcciones futuras se propone el desarrollo de benchmarks estandarizados para conocimiento operativo, estudios comparativos entre paradigmas de aprendizaje y la integración de los criterios propuestos en marcos formales de auditoría y certificación de sistemas de IA.

REFERENCIAS

- [1] M. Baltieri, C. L. Buckley y M. McGann, “Embodied artificial intelligence through distributed adaptive control,” *Artificial Life*, vol. 26, no. 4, pp. 1–25, 2020.
- [2] E. M. Bender, T. Gebru, A. McMillan-Major y S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?,” *Proc. ACM FAccT*, pp. 610–623, 2021.
- [3] D. J. Chalmers, *Reality+*, New York, NY, USA: W. W. Norton, 2022.
- [4] F. Dretske, *Knowledge and the Flow of Information*, Cambridge, MA, USA: MIT Press, 1981.
- [5] L. Floridi, *The Philosophy of Information*, Oxford, UK: Oxford Univ. Press, 2011.
- [6] E. Gettier, “Is justified true belief knowledge?,” *Analysis*, vol. 23, no. 6, pp. 121–123, 1963.
- [7] B. Lake et al., “Building machines that learn and think like people,” *Behav. Brain Sci.*, vol. 40, e253, 2017.
- [8] Y. LeCun, “A path towards autonomous machine intelligence,” *OpenReview*, 2022.
- [9] M. Lungarella, G. Metta, R. Pfeifer y G. Sandini, “Developmental robotics: A survey,” *Connection Science*, vol. 15, no. 4, pp. 151–190, 2003.
- [10] G. Marcus y E. Davis, *Rebooting AI*, New York, NY, USA: Pantheon, 2019.
- [11] T. Metzinger, “Artificial consciousness: An update,” *Journal of Consciousness Studies*, vol. 30, no. 1–2, pp. 77–102, 2023.

- [12] M. Mitchell, *Artificial Intelligence: A Guide for Thinking Humans*, New York, NY, USA: Farrar, Straus and Giroux, 2019.
- [13] J. Pearl y D. Mackenzie, *The Book of Why*, New York, NY, USA: Basic Books, 2018.
- [14] R. Pfeifer y J. Bongard, *How the Body Shapes the Way We Think*, Cambridge, MA, USA: MIT Press, 2006.
- [15] H. Putnam, "The meaning of 'meaning'," *Minnesota Studies in the Philosophy of Science*, vol. 7, pp. 131–193, 1975.
- [16] S. Reed et al., "A generalist agent," *Trans. Mach. Learn. Res.*, 2022.
- [17] A. Seth, *Being You*, London, UK: Faber & Faber, 2021.
- [18] F. Varela, E. Thompson y E. Rosch, *The Embodied Mind*, Cambridge, MA, USA: MIT Press, 1991.